

# A New Use of Doppler Spectrum for Action Recognition with the Help of Optical Flow



Meropi Pavlidou and George Zioutas

**Abstract** In this work, we present two new procedures for activity recognition that are based on the Fourier frequencies that are generated when the optical flow values of successive frames of video are processed simultaneously. In the first algorithm, we correlate these 2D Doppler Fourier spectra with the mean spectra of each activity class. These correlation vectors, which include only 30 features in number, are categorized using a reduced robust SVM classification model. This first procedure is of low computational cost for action recognition tasks for numerable activity classes. For large numbers of activity classes, we propose a new method of aggregated weighted spectra of optical flow values across the whole video. The above-mentioned Fourier spectra are concatenated with a short vector representing the distributions of the moving edges. These methods are insensitive to the presence of background as well as to the positions of the subjects and their shapes and can encode the information of a part or of the whole of a video into relatively short vectors. The results of the two procedures seem to be competitive to state-of-the-art action recognition methods when tested on the KTH Royal Institute Database and on the UCF101 Database for action recognition tasks.

**Keywords** Action recognition · Doppler frequencies · Spectral density estimation · Kalman Filter · Optical flow · SVM classification · Principal component analysis

## 1 Introduction

Activity recognition plays an important role for many different applications such as health care, human–computer interaction or social sciences. Action recognition algorithms are based either on global features or on local features. Spatiotemporal feature points based on local movements aim at robustness to pose, image clutter, occlusion

---

M. Pavlidou (✉) · G. Zioutas  
Aristotle University of Thessaloniki, Thessaloniki, Greece  
e-mail: [mepa@auth.gr](mailto:mepa@auth.gr)

G. Zioutas  
e-mail: [zioutas@eng.auth.gr](mailto:zioutas@eng.auth.gr)

© Springer Nature Singapore Pte Ltd. 2020  
X.-S. Yang et al. (eds.), *Fourth International Congress on Information and Communication Technology*, Advances in Intelligent Systems and Computing 1041, [https://doi.org/10.1007/978-981-15-0637-6\\_35](https://doi.org/10.1007/978-981-15-0637-6_35)

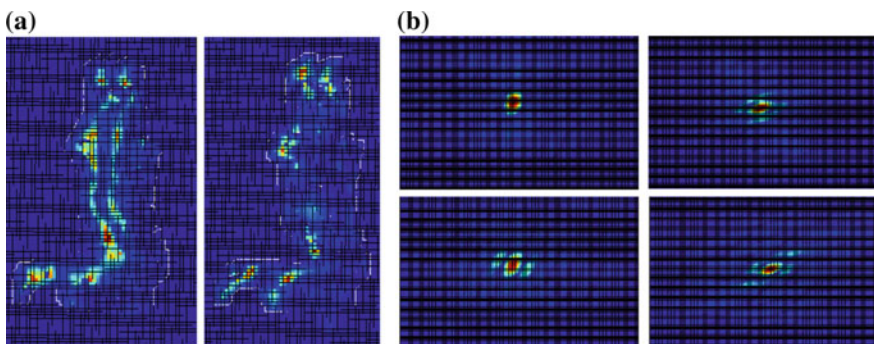
and object variation in [1]. In [2], dynamic time warping and posterior probability model the gait of the subjects. The work of Efros et al. [3] proposes a spatiotemporal descriptor based on optical flow estimation. Probabilistic latent semantic analysis and latent dirichlet allocation in [4] estimates the probability distributions of the spatial–temporal words. Neural networks for computational intelligence applications [5–7] are applied to the frames directly or to trajectories extracted by multiple frames [8].

Many of the algorithms that are currently used for activity recognition include large vectors of data and are sensitive to the different frame sizes and the resolution of the input videos. In an effort to increase the amount of information coming from multiple frames, we decided to combine the Doppler frequencies that are generated from the optical flows of the moving edges of multiple frames.

In this paper, we take advantage of the Doppler frequencies that are generated when we concatenate the 2D optical flow values of the pixels of the moving edges. The vectors that are produced from our two methods, two-dimensional correlation coefficient Doppler spectroscopy descriptor (CCDS) and weighted Doppler spectroscopy descriptor (WDS), are relatively short, may include information from a part of the movement video or from the whole video and are insensitive to noise, background or brightness.

## 2 Two-Dimensional Correlation Coefficient Doppler Spectroscopy Descriptor (CCDS)

In the CCDS algorithm, the first step is to identify the moving edges of two consecutive frames with the help of the Kalman filter. In the second step, the Horn–Schunck [9] optical flow values of those edges are computed, as in Fig. 1a. Since those optical flow values regard two consecutive video frames and are present in the same matrix,



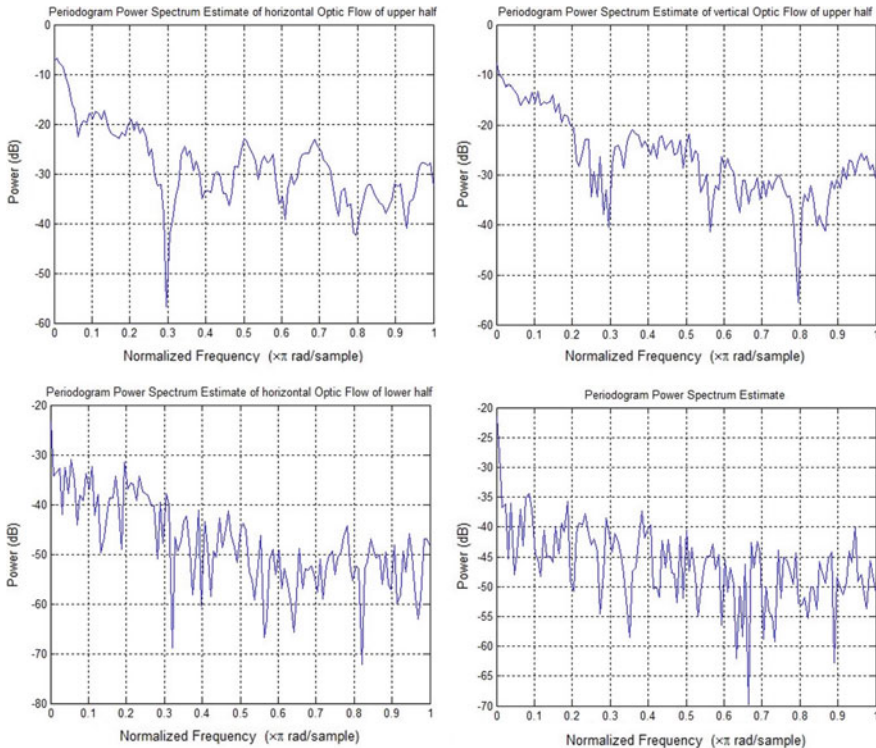
**Fig. 1** **a** Horizontal and vertical optical. **b** Two-dimensional spectral density of the flow velocities present from two con-upper-half and the lower half of the jogging man seductive frames of a person jogging frame

Doppler frequencies are generated. Those Doppler frequencies are detected when we estimate the two-dimensional Fourier spectrum of this matrix as shown in Fig. 1b.

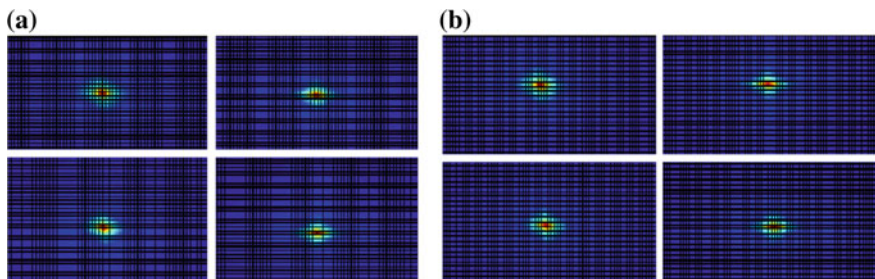
The discrete Fourier transform  $Y$  of an  $m$ -by- $n$  matrix  $X$ :

$$Y_{p+1,q+1} = \sum_{j=0}^{m-1} \sum_{k=0}^{n-1} w_m^{jp} w_n^{kq} X_{j+1,k+1} \tag{1}$$

For better classification results, and inspired from [10] where halves of the frames are used in computations, we divide the optical flow values in half into those belonging to the upper half and the lower half, and also we divide them into the vertical and the horizontal optical flow values. Consequently, we now can compute the Doppler spectra of four two-dimensional matrices as we can see in Fig. 2. In the next step, we calculate the values of the final vector prior the classification. These values are the 2D correlations between the four spectra of each frame, as in Fig. 2, and the average spectra of each class of data are, for example, shown in Fig. 3a, b.



**Fig. 2** Periodograms of horizontal and vertical values of optical flows for the upper and lower half of a boxing video



**Fig. 3** **a** Averaged spectra of the Doppler frequencies of the horizontal (a), (c) and vertical (b), (d) optical flow values of the upper (a), (b) and lower (c), (d) halves of class Boxing of the KTH Activity Database [11]. **b** Averaged spectra of the Doppler frequencies of the horizontal (a), (c) and vertical (b), (d) optical flow values of the upper (a), (b) and lower (c), (d) halves of class walking of the KTH activity database [11]

Two-dimensional correlation analysis is a mathematical technique that is used to study changes in measured signals. Let us assume that we have in our disposition two spectral densities of two two-dimensional optical flow velocities,  $F^k$  and  $F^M$ .  $F^k$  is the spectral density of the optical values of frame  $k$  and  $F^M$  The two-dimensional correlation is the mean spectral density of the optical values of all the frames of class  $M$ , for example, of the activity of jogging. Then, the two-dimensional correlation,  $\text{Corr}_{2D}$ , between  $F^k$  and  $F^M$  is calculated as:

$$\text{Corr}_{2D} = \frac{\sum_m \sum_n (F_{mn}^k - \overline{F_{mn}^k})(F_{mn}M - \overline{F_{mn}M})}{\sum_m \sum_n (F_{mn}^k - \overline{F_{mn}^k}) \sum_m \sum_n (F_{mn}M - \overline{F_{mn}M})} \quad (2)$$

$\text{Corr}_{2D}$  is used as a metric of similarity of the spectral density of each frame with the mean spectral density of the optical flow velocities of all the frames in a specific class. If we correlate the spectral densities for all the frames and all the classes, for the horizontal and vertical velocities, we end up with a vector of correlation coefficients of size  $2N$  for each frame.  $N$  is the number of the classes, and therefore, the number of the means of the spectral densities is 2 because we have the densities for the horizontal and vertical velocities separately. Concluding the CCDS method so far, the proposed algorithm for the estimation of the correlated spectral density activity descriptors follows the subsequent procedure:

- Use Kalman filter to detect motion in each video and choose only those frames that are positive for movement detection and the pixels in each frame that are positive for movement.
- Convert each frame from RGB to greyscale.
- Compute the optical flow values using Horn–Schunck algorithm for the pixels of interest.
- Estimate the absolute values of the optical flow as we are interested in the velocity of the edges and not the direction of the movement.

- Combine the optical flow values of the moving edges of two consecutive frames by adding them, creating a 2D matrix that holds the information of pixels moving relatively to each other with specific velocities and thus generating Doppler frequencies.
- Estimate the spectral densities of the vertical and the horizontal concatenated optical flow values in the upper half and the lower half of the image.
- Correlate those spectrum densities with the mean spectrum densities of each activity class.

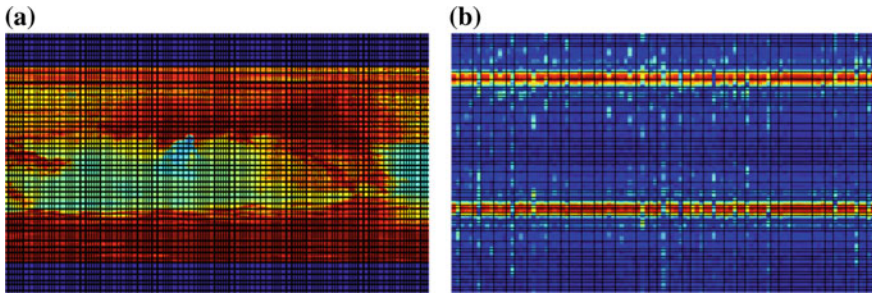
At this point, we have in our disposition a vector for each frame of every video. This vector holds the concatenated spectral densities of the optical flow values, horizontal and vertical, of the upper half and lower half, which were detected by the Kalman filter. The next step is to apply the classification method for each vector so that we can recognize human activity.

### 3 Weighted Doppler Spectroscopy Descriptor (WDS)

We already described how, as objects move across the frames, the pixels that show nonzero velocities create moving frames. In Fig. 5, the horizontal and vertical velocities of two consecutive frames of a person jogging are present. However, in our experiments we noticed that when the number of classes is increased and along with them the number of frames and the range of the activities, then CCDS which draws information based on only two consecutive frames is easier to misclassify. The CCDS information of the two frames is the partial information of a movement that may regard seconds or even a fraction of a second and may not be able to represent the nature of the action in a video.

At this point, the need for an algorithm that efficiently combines the knowledge of all the frames taking into consideration the direction of the evolution of the action seems as the next step that we need to take. So how about we gather the information not from just two frames, because it is too scarce, but from the whole video, all of its frames? How about having the vertical and the horizontal velocities of all the frames present in two matrices? In this way, we can produce a single matrix that holds the Doppler spectrum information of the optical flow values of all the frames of the video and represents the whole action and not just a piece of the action like the information provided by CCDS that regards only two frames. Below we present an example of the spectra of all the video frames of a surfing video that belongs to the UCF101 Dataset [12]. The respective Doppler spectra of the upper and the lower halves of the horizontal and vertical velocities would be those of Fig. 4b.

In Fig. 4b, we can see the spectra of the velocities, horizontal and vertical, of moving pixels of all frames of class Surfing of the UCF101 Dataset, of the upper halves and lower halves. The velocities of all the frames are summed up. This unweighted summation provides no information on the evolution of the action depicted in the video. Therefore, a weight is multiplied with the velocities of the moving edges of



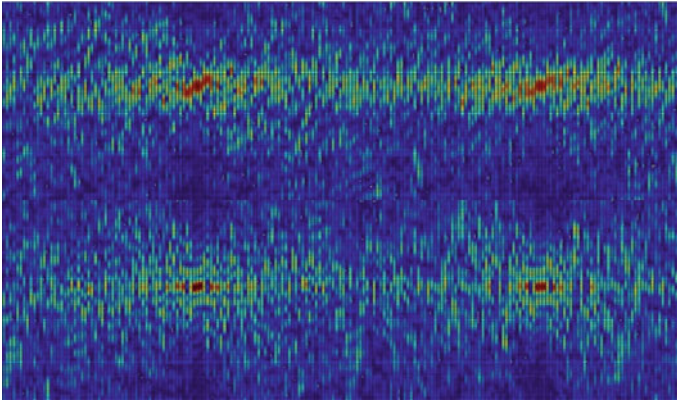
**Fig. 4** **a** Moving edges of all frames of class Surfing of the UCF101 Dataset combined in one matrix. **b** Spectra of the velocities, horizontal and vertical, of moving pixels of all frames of class Surfing of the UCF101 Dataset, of the upper halves and lower halves. These are the spectra of the summed up velocities of every frame

every frame. For the first frame, the weight is the number of the frames of the video. For the second frame, this number is reduced by one and so on until the weight for the last video becomes equal to one. Finally, these weighted velocities are summed up and normalized, and their spectra are calculated for the upper half of the frames, the lower half, the horizontal and the vertical velocities. The matrix of the Spectra of every video is reshaped into a vector. We can see distinct shapes that differ from one another in terms of shape, inclination, dispersion and intensity. Moreover, for the sake of computational complexity, we have reduced the resolution of the spectra during the validation step down to a level that produces reliable results without the need of extremely large matrices for the spectra. Moreover, in order to clean each class of the Training subset, we apply the robust principal component analysis (Robust PCA) method [13–16] to clean the data of each class of the Training matrix, thus producing the final clean vector for each video without outliers. Perform one-against-all (OAO) SVM classification, and assign each data point to the class for which it gathers the most scores, above 50%. This vector will be concatenated with the reshaped spectra vector, and this final vector will comprise our new action recognition descriptor.

Concluding the WDS method, the proposed algorithm for the estimation of the weighted Doppler spectral density activity descriptors follows the subsequent procedure:

- Use Kalman filter to detect motion in each video, and choose only those frames that are positive for movement detection and the pixels in each frame that are positive for movement.
- Convert each frame from RGB to greyscale.
- Compute the optical flow values using Horn–Schunck algorithm for the pixels of interest.
- Estimate the absolute values of the optical flow as we are interested in the velocity of the edges and not the direction of the movement.
- Combine the optical flow values of the moving edges of all consecutive frames by adding them, with decreasing weights that decline from the number of frames to





**Fig. 5** Spectra of the velocities, horizontal and vertical, of moving pixels of all frames of class Surfing of the UCF101 [12] dataset, of the upper halves and lower halves. These are the spectra of the weighted summed up velocities of every frame

- 1, creating a 2D matrix that holds the information of edges moving relatively to each other with specific velocities and thus generating Doppler frequencies.
- Estimate the spectral densities of the vertical and the horizontal concatenated optical flow values in the upper half and the lower half of the image. Concatenate the matrices and reshape them into one vector for each video.
- Apply the robust PCA method to clean the data of each class of the Training matrix, thus producing the final clean vector for each video without outliers.
- Perform OAO linear SVM classification, and assign each data point to the class for which it gathers the most scores, above 50% .

## 4 Experimental Results

### 4.1 *KTH Royal Institute of Technology Human Activity Database*

We tested our model on numerous human activity datasets. The first results come from KTH Royal Institute of Technology [11]. This video database containing six types of human actions, walking, jogging, running, boxing, hand waving and hand clapping, performed several times by 25 subjects in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors as illustrated below. Currently, the database contains 2391 sequences. All sequences were taken over homogeneous backgrounds with a static camera with 25 fps frame rate. The sequences were downsampled to the spatial resolution of  $160 \times 120$  pixels and have a length of four seconds in average.

**Table 1** Classical SVM classification on periodogram of quantized optical flow values of KTH Royal Institute Database

SVM	Yeo [17]	Sch. [11]	Dol. [1]	Nieb. [4]	Hist. [10]	Per.
handcl.	0.89	0.60	0.76	0.93	0.78	0.9015
handw.	0.95	0.74	0.88	0.77	0.75	0.8673
box.	0.82	0.97	0.93	1.00	0.86	0.9059
jog.	0.58	0.60	0.13	0.52	0.92	0.8976
run.	0.91	0.55	0.84	0.88	0.92	0.9344
walk.	1.00	0.84	0.89	0.79	0.89	0.9201

We notice that in the cases of handclapping, hand waving and boxing, Nieb. [4] outperforms the periodogram descriptors by 2%. However, this method is outperformed in the cases of jogging, running and walking and specifically has a recognition score of 13% in the case of jogging. Sch. [11] method is outperformed by 2–38% in every case except boxing where the recognition rate is 7% higher. Yeo [17] is also outperformed in handclapping, boxing, jogging and running by 2–31% while it outperforms the periodogram descriptors by 8% in walking and running. Finally, Hist. [10] method, which does not display recognition rates lower than 75% and shows high recognition efficiency, is outperformed from 1.44 to 12% in every case except the boxing case where the result is better by 2.3% compared to the periodogram algorithm. Concluding, we can say that the new method outperforms the up-to-date action recognition methods and is characterized by steady behaviour across the six different actions of handclapping, hand waving, boxing, jogging, running and walking (Table 1).

## 4.2 University of Central Florida, UCF101, Human Activity Database

UCF101 [12] is an action recognition dataset of realistic action videos, collected from YouTube, having 101 action categories. UCF101 includes 13,320 videos from 101 action categories, providing with a large diversity in terms of actions and with the presence of large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, as well as illumination conditions and is considered to be one of the most challenging dataset to date. The action categories in UCF101 can be divided into five types: human–object interaction, body-motion only, human–human interaction, playing musical instruments and sports. Specifically, the action categories for UCF101 dataset are: Apply Eye Makeup, Apply Lipstick, Archery, Baby Crawling, Balance Beam, Band Marching, Baseball Pitch, Basketball Dunk, Basketball, Bench Press, Biking, Billiards, Blow Dry Hair, Blowing Candles, Body Weight Squats, Bowling, Boxing Punching Bag, Boxing Speed Bag, Breaststroke, Brushing Teeth, Clean and Jerk, Cliff Diving, Cricket Bowling,



Cricket Shot, Cutting in Kitchen, Diving, Drumming, Fencing, Field Hockey Penalty, Floor Gymnastics, Frisbee Catch, Front Crawl, Golf Swing, Haircut, Hammer Throw, Hammering, Handstand Pushups, Handstand Walking, Head Massage, High Jump, Horse Race, Horse Riding, Hula Hoop, Ice Dancing, Javelin Throw, Juggling Balls, Jump Rope, Jumping Jack, Kayaking, Knitting, Long Jump, Lunges, Military Parade, Mixing Batter, Mopping Floor, Nun chucks, Parallel Bars, Pizza Tossing, Playing Cello, Playing Daf, Playing Dhol, Playing Flute, Playing Guitar, Playing Piano, Playing Sitar, Playing Tabla, Playing Violin, Pole Vault, Pommel Horse, Pull Ups, Punch, Push Ups, Rafting, Rock Climbing Indoor, Rope Climbing, Rowing, Salsa Spins, Shaving Beard, Shot-put, Skate Boarding, Skiing, Ski jet, Sky Diving, Soccer Juggling, Soccer Penalty, Still Rings, Sumo Wrestling, Surfing, Swing, Table Tennis Shot, Tai Chi, Tennis Swing, Throw Discus, Trampoline Jumping, Typing, Uneven Bars, Volleyball Spiking, Walking with a dog, Wall Pushups, Writing On Board and Yo Yo. The downloaded data for the UCF101 dataset did not include any videos for the classes. For the purpose of classifying UCF101 activity videos, up-to-date classification methods have been applied. In [18], a ConvNet model which forms a temporal recognition stream is formed. This model is formed by stacking optical flow displacement fields between several consecutive frames which facilitate the description of the motion between video frames and make the recognition easier, as the network does not need to estimate motion implicitly. The Temporal ConvNets achieved a remarkable 81.2% recognition rate on the UCF101 database. [19] applied a flexible and efficient Temporal Segment ConvNet and succeeded a 94.2% recognition rate on the UCF101. We also inserted the WDS vectors in Sect. 5 Accelerated SVM method performs one-against-one (OAO) classification for each of the 101 classes of the UCF101. Each data sample is assigned to the class that gets the higher score among the OAO classification tasks. If the sample is correctly assigned, then it is regarded as being correctly recognized. The recognition rates of the WDS algorithm are presented in Table 2. The training and testing vectors can be provided through an e-mail to mepa@auth.com.

## 5 Conclusions

In this research, two new action recognition methods were developed for action recognition that are based on a new use for the Doppler frequencies. We combined the optical flows of the moving pixels of two subsequent frames for the CCDS method and for all the frames with the help of weights for the WDS method. The presence of Optical flow values in 2D matrices that express the evolution of the movement and generate Doppler frequencies. These frequencies are subsequently used for Activity Recognition with the help of Support Vector Machines. In the case of the WDS, a robust accelerated SVM method was used since the number of videos for the first experiment was limited and the number of the training and testing vectors is also limited. When we proceeded to the second experiment with the UCF101 dataset, the number of videos and therefore frames increased significantly. This fact

**Table 2** Recognition rates using the WDS method on the UCF101 Database

Apply Eye Makeup	92.68	Biking	86.4864	Clean and Jerk	(Missing data)
Apply Lipstick	78.1250	Billiards	100	Cliff Diving	70.1244
Archery	90	Blow Dry Hair	88.8888	Cricket Bowling	97.4358
Baby Crawling	91.8919	Blowing Candles	(Missing data)	Cricket Shot	91.4893
Balance Beam	86.6667	Body Weight Squats	74.19354	Cutting in Kitchen	76.6666
Band Marching	100	Bowling	100	Diving	88.8888
Baseball Pitch	100	Boxing Punching Bag	73.20354	Drumming	93.3333
Basketball Dunk	78.378	Boxing Speed Bag	100	Fencing	35.4838
Basketball	100	Breast Stroke	85.7142	Field Hockey Penalty	80
Bench Press	46.666	Brushing Teeth	94.444	Floor Gymnastics	94.2857
Frisbee Catch	97.1428	Horse Race	91.1764	Long Jump	100
Front Crawl	100	Horse Riding	82.6086	Lunges	71.4285
Golf Swing	65	Hula Hoop	62.8571	Military Parade	82.857
Haircut	72.2222	Ice Dancing	97.7272	Mixing Batter	92.1052
Hammer Throw	82.0512	Javelin Throw	84.3750	Mopping Floor	96.6666
Hammering	90.4761	Juggling Balls	70.5882	Nun chucks	86.4864
Handstand Pushups	83.3333	Jump Rope	76.4705	Parallel Bars	96.8750
Handstand Walking	93.5483	Jumping Jack	82.5000	Pizza Tossing	87.0967
Head Massage	80.4878	Kayaking	97.4358	Playing Cello	93.4782
High Jump	91.1764	Knitting	88.2352	Playing Daf	90.4761
Playing Dhol	65.2173	Punch	98.9010	Skiing	89.4736
Playing Flute	67.4418	Push Ups	50	Ski jet	92.8571
Playing Guitar	84.4444	Rafting	80.6451	Sky Diving	80

(continued)

**Table 2** (continued)

Playing Piano	89.6551	Rock Climbing Indoor	77.5000	Soccer Juggling	78.0487
Playing Sitar	88.6363	Rope Climbing	81.8181	Soccer Penalty	94.7368
Playing Tabla	80.6451	Rowing	81.5789	Still Rings	93.5483
Playing Violin	89.2857	Salsa Spin	91.8918	Sumo Wrestling	84.3750
Pole Vault	83.3333	Shaving Beard	88.8888	Surfing	91.4285
Pommel Horse	85.2941	Shot-put	95	Swing	89.4736
Pull Ups	85.71428	Skate Boarding	90.9090	Table Tennis Shot	92.3076
Tai Chi	64.2857	Tennis swing	95.6521	Throw Discus	88.8888
Trampoline Jumping	87.8787	Typing	100	Uneven Bars	86.2068
Volleyball Spiking	71.8750	Walking with dog	85.2941	Wall Pushups	77.7777
Writing on Board	85.7142	Yo Yo	72.9729		

along with the significant increase in the number of the classes, from 6 classes of the first experiment to 101 classes, created the need for a more robust method that would save computational time and cost. This is how we multiplied each optical flow frame values with weights, and we added them in order to get one single vector for each video. The vector is the reshaping of a 2D spectrum of size  $102 \times 102$  which results in a vector of 10,404 elements. If we want even better resolution, we will probably increase the resolution of the 2D Fourier transform which will generate larger vectors that will include more frequencies. The choice of our parameters for the experiments was performed by using one-third of the videos from every one of the 101 classes for validation purposes. The data after the final processing and checks are available by sending an e-mail to the authors.

**Acknowledgements** The KTH Royal Institute of Technology video database [11] that was used is publicly available for non-commercial use.

The UCF101, Human Activity Database [12] is freely available here: <https://www.crcv.ucf.edu/data/UCF101.php>.

## References

1. P. Dollar, V. Rabaud, G. Cottrel, S. Belongie, Behavior “recognition via spatiotemporal features”, in *Proceedings 2nd Joint IEEE International Workshop on VS-PETS, Beijing* (IEEE Computer Society Press, Los Alamitos, 2005), pp. 65–72
2. L. Wang et al., Fusion of static and dynamic body biometrics for gait recognition. *IEEE Trans. Circ. Syst. Video Technol.* **14**(2), 149–158 (2004)
3. A.A. Efros, et al., Recognizing action at a distance, in *ICCV*, vol. 3 (2003)
4. J.C. Niebles, H. Wang, L. Fei-fei: Unsupervised learning of human action categories using spatial-temporal words, in *BMVC* (2006)
5. L.P. Wang, X.J. Fu, *Data Mining with Computational Intelligence* (Springer, Berlin, 2005)
6. X.J. Fu, L.P. Wang, Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance. *IEEE Trans. Syst. Man Cybern Part B Cybern* **33**(3), 399–409 (2003)
7. L.P. Wang, On competitive learning. *IEEE Trans. Neural Networks* **8**(5), 1214–1217 (1997)
8. V. Luong, L.P. Wang, G. Xiao, Deep networks with trajectory for action recognition in videos, in *The 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES 2014)*, Singapore, 10–12th Nov 2014
9. B.K.P. Horn, B.G. Schunck, Determining optical flow. *Artif. Intell.* **17**(1–3), 185–203 (1981)
10. S. Danafar, N. Gheissari, Action recognition for surveillance applications using optic flow and SVM, in *Asian Conference on Computer Vision* (Springer, Berlin, 2007)
11. C. Schudt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in *ICPR* (2004), pp. 32–36
12. K. Soomro, A.R. Zamir and M. Shah, UCF101: a dataset of 101 human action classes from videos in the wild. CRCV-TR-12-01, November, 2012.
13. M. Hubert, P.J. Rousseeuw, K. Vanden Branden, ROBPCA: a new approach to robust principal components analysis. *Technometrics* **47**, 64–79 (2005)
14. M. Hubert, S. Engelen, Fast cross-validation of high-breakdown resampling algorithms for PCA. *Comput. Stat. Data Anal.* **51**, 5013–5024 (2007)
15. M. Hubert, P.J. Rousseeuw, T. Verdonck, Robust PCA for skewed data and its outlier map. *Comput. Stat. Data Anal.* **53**, 2264–2274 (2009)
16. S. Serneels, T. Verdonck, Principal component analysis for data containing outliers and missing elements. *Comput. Stat. Data Anal.* **52**, 1712–1727 (2008)
17. P. Ahammad, C. Yeo, S.S. Sastry, K. Ramchandran, Compressed domain real-time action recognition, MMSP, in *Proceedings of 8th IEEE Workshop on Multimedia Signal Processing* (IEEE Computer Society Press, Los Alamitos, 2006) pp. 33–36
18. K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in *Advances in Neural Information Processing Systems* (2014)
19. L. Wang, et al., Temporal segment networks: towards good practices for deep action recognition, in *European Conference on Computer Vision* (Springer International Publishing, 2016)